

# Can Shallow Predicate Argument Structures Determine Entailment?

Alina Andreevskaia, Zhuoyan Li and Sabine Bergler

## Abstract

The CLaC Lab’s system for the PASCAL RTE challenge explores the potential of simple general heuristics and a knowledge-poor approach for recognising paraphrases, using NP coreference, NP chunking, and two parsers (RASP and Link) to produce Predicate Argument Structures (PAS) for each of the pair components. WordNet lexical chains and a few specialised heuristics are used to establish semantic similarity between corresponding components of the PAS from the pair. We discuss the results and potential of this approach.

## 1 Introduction

Establishing entailment relationships between two statements is important for many NLP tasks (Szpektor et al., 2004) and the problem has attracted considerable interest in the research community. Most current work relies on the analysis of corpora - single or parallel - using machine learning and statistical methods (Lin and Pantel, 2001), (Chklovski and Pantel, 2004), (Dagan and Glickman, 2004), (Shinyama and Sekine, 2004), (Barzilay and Lee, 2003) to induce entailment-specific knowledge. In contrast, we approach the textual entailment problem using general mechanisms and strategies based uniquely on simplified predicate argument structure (PAS) and lexical chains (built using WordNet (WN) (Fellbaum, 1998)). This paper de-

scribes the results we achieved with this simple approach and suggests extensions to improve system performance.

## 2 System overview

The CLaC Lab’s system for the PASCAL textual entailment challenge is based on systems our laboratory developed for text summarization. The environment is implemented in the GATE architecture (Cunningham et al., 2002) and provides tagging, NP chunking, and knowledge-poor fuzzy NP coreference resolution (Bergler et al., 2003), (Bergler et al., 2004), (Witte and Bergler, 2003). The flexible GATE architecture allows for the creation of modular components that can be used in different combinations depending on the task. For the purposes of the textual entailment resolution we extended the coreference system to incorporate verb groups, added full parsing, and included a few specialized heuristics for particular problems that were encountered in the PASCAL RTE challenge development set.

### 2.1 Main strategy

Two main types of information were used to assess the relatedness between the two parts of the pair: PAS and lexical chains. We use simplified shallow PASs that cover only the verb, its subject and object (if there was one) as in Figure 1.

PASs were extracted using the results of two parsers - the Link parser (Sleator and Temperley, 1993) and RASP parser (Briscoe and Carroll, 2002). One of these two parsers can be set as default, the second to be used only when the default parser doesn’t produce a parse. If

<p><b>Original Sentence:</b>  &lt;h&gt;Two-thirds of the Scottish police force will be deployed at the happening.&lt;/h&gt;</p> <p><b>Constructed PAS:</b>  &lt;s:[Two-thirds of the Scottish police force] v:[will be deployed] a:[&lt;p:[at] a:[the happening]&gt;]&gt;</p>
---

Figure 1: Predicate argument structure

both parsers are given equal priority the system chooses for each sentence the parser that produces more PASs. Lexical chains were built using WN synsets. Different thresholds were tested. The smaller values mean closer relationships, 0 being the distance between members of the same synset.

### Algorithm CLaC PASCAL

(\* **true**: entailment detected, **false** otherwise )

1. Use the coreference resolution system to produce coreference chains both for  $t$  and  $h$  separately and for the pair as a unit
2. **for** each pair
3.     **for** each sentence
4.         Extract Noun Phrases and Verb Groups
5.         Select a parse among parses from two parsers with weighted scheme
6.         Determine the PAS based on the parsing, NP chunking and verb grouping results
7.     *Apply cardinality filter*
8.         **for** each numeric value from  $h$
9.             **if** there is no corresponding cardinality value in  $t$
10.                 **then return false**
11.     *Apply Predicate Argument Structure comparison*
12.         Transform passive constructions into active ones
13.         **for** each PAS pair
14.             Compute WN distance for verbs in  $t$  and  $h$
15.             **if** WN distance  $\leq$  threshold
16.                 **if** both PASs are in *comparable structures*<sup>1</sup>

<sup>1</sup>*comparable structure* means they both have sub-

17.                     **if** there is coreference between corresponding parts<sup>2</sup>
18.                     **then return true**
19.     *Apply Be-Heuristic*
20.         **if**  $h$  contains the pattern “X is Y” **and**  $X \in h$  **and**  $X' \in t$  **and**  $\{X, X'\}$  belong to the same inter-sentence coreference chain **and**  $Y \in h$  **and**  $Y' \in t$  **and**  $\{Y, Y'\}$  belong to the same inter-sentence coreference chain **and**  $X'$  corefers with  $Y'$
21.                     **then return true**
22.     **return false**

The algorithm favors precision over recall, therefore all entailment values are set to FALSE unless the system finds compelling evidence to the contrary.

Analysis of the development data allowed us also to develop some additional heuristics to handle specific cases. For example, we have implemented a *be-heuristic* for  $h$ -sentences of type “X is Y” that uses coreference chains in  $t$  and between  $t$  and  $h$  to decide whether X is Y given the data in  $t$ . The development data contains many examples of this kind in the QA task, but the phenomenon was less frequent in the test data. Another heuristic consists in comparing numbers in two parts of the pair to ensure that cases like pair 768 (Figure 2) from the development set do not produce false positives. This heuristic is applied as an initial filter before coreference chains are built.

<t>A small bronze bust of Spencer Tracy sold for £174,000.</t>  
<h>A small bronze bust of Spencer Tracy made £180,447.</h>

Figure 2: Cardinality filtering example

## 2.2 Results

We submitted two runs, Table 1 presents the results of both runs, where RASP was the main parser and the Link parser used only as backup, RUN1 used a WN distance threshold of 1.

ject(s) and/or argument(s)

<sup>2</sup>e.g. subjects and/or arguments of the two PASs being compared

The second run used a WN threshold of 3. Our

Task	RUN1			
	P	R	A	Cws
All	0.57	0.15	0.52	0.51
CD	0.89	0.32	0.64	0.64
IE	0.56	0.08	0.51	0.55
MT	0.40	0.10	0.47	0.43
QA	0.23	0.04	0.45	0.47
RC	0.52	0.17	0.51	0.48
PP	0.50	0.28	0.50	0.54
IR	0.62	0.11	0.52	0.49
Task	RUN2			
	P	R	A	Cws
All	0.55	0.18	0.52	0.52
CD	0.81	0.34	0.63	0.63
IE	0.64	0.12	0.52	0.57
MT	0.37	0.10	0.47	0.43
QA	0.31	0.08	0.45	0.49
RC	0.44	0.17	0.48	0.47
PP	0.50	0.36	0.50	0.56
IR	0.64	0.16	0.53	0.49

Table 1: Results over the different categories

conservative strategy lead to a low number of true-positives: 72 true-positives of 400 in the gold standard in RUN2.

### 3 Analysis and observation

Our main interest in participating in the PAS-CAL RTE challenge was to experiment with simple general purpose tools such as a coreference resolution system and a parser for textual entailment recognition.

The performance of our system is low, as expected, but comparable to the results shown by other systems. Most correct TRUE assignments occur when PASs are properly extracted and there is considerable similarity between PASs of  $t$  and  $h$ . This explains also the difference between our results for different tasks. CD, for instance, gave the highest precision (0.89 in Run1, 0.92 when WN distance=1 and parsers have equal weight), since pairs are mostly made up of sentences of similar structure (Figure 3) while QA consistently gave the worst results (precision below 0.30 and accuracy below 0.50), since

it includes answers derived from statements of a totally different structure (Figure 3). In general, most of our false negatives are due to not recognising similarity between two syntactically different sentences. More sophisticated PASs that include additional constituents, such as adjuncts, and specialized heuristics geared towards frequent syntactic patterns in the data, as we did for the *be-heuristic* would address these issues.

<t>In terms of music, the National Philharmonic Orchestra draws large crowds.</t>  
<h>The National Philharmonic orchestra draws large crowds.</h>

Figure 3: Correctly processed pair

<t>Working with fellow Canadians Charles Best and James Collip, Banting determined that insulin was the key to treating diabetes.</t>  
<h>Banting conducted research of diabetes.</h>

Figure 4: False negative

The parser influences the system’s performance (Table 2), best results are obtained when the choice between the Link and RASP parsers depends on the number of PASs produced, thus increasing the chances to find comparable PASs in  $t$  and  $h$  parts of the pair. Making PASs more complex by including prepositional phrases and adjuncts can eliminate such false positives as in Figure 5.

<t>The 69-page report is also the first major product of the Betsy Lehman Center for Patient Safety and Medical Error Reduction.</t>  
<h>The 69-page report is the first major product of medical errors.</h>

Figure 5: False positive

Table 3 illustrates the influence of the WN distance threshold when both parsers have equal priority and the one producing more parses is preferred. Increasing the WN distance threshold leads to increased recall but reduced precision since more PASs are considered semantically related.

Setting	P	R	A	Cws
Equal priority	.59	.13	.52	.52
RASP/Link	.55	.13	.52	.51
Link/RASP	.58	.12	.52	.51

Table 2: Post-competition runs, WN distance=0

	P	R	A	Cws
WD= 0	.59	.13	.52	.52
WD= 1	.56	.15	.52	.52
WD= 2	.55	.16	.51	.52
WD= 3	.52	.18	.51	.52

Table 3: Influence of WN distance threshold (WD)

## 4 Conclusion

The PASCAL RTE challenge gave us an opportunity to create and test a system that we consider as a baseline system for our future work on event coreference and analysis of comparable documents. Our simple approach based on basic PASs and coreference resolution produced the precision slightly below 0.6 (up to 0.92 for the CD task). At the same time, the recall was fairly low - 0.18 (best value being 0.36 for PP task). These numbers can be improved by applying more sophisticated PASs and by creating additional heuristics to deal with specific patterns.

## References

- Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In *Proceedings of HLT-NAACL 2003*, pages 16–23.
- Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, and Frank Rudzicz. 2003. Using Knowledge-poor Coreference Resolution for Text Summarization. In *Workshop on Text Summarization*, Document Understanding Conference (DUC), Edmonton, Canada, May 31–June 1. NIST.
- Sabine Bergler, René Witte, Michelle Khalife, Zhuoyan Li, Yunyu Chen, Monia Doandes, and Alina Andreevskaia. 2004. Multi-ERSS and ERSS 2004. In *Workshop on Text Summarization*, Document Understanding Conference (DUC), Boston, MA, May 6–7. NIST.
- E. Briscoe and J. Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands, May 2002.
- Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 33–40, Barcelona, Spain, July. Association for Computational Linguistics.
- H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.
- Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability. In *Learning Methods for Text Understanding and Mining Workshop*. January.
- Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical database*. MIT Press.
- Dekang Lin and Patrick Pantel. 2001. DIRT-discovery of inference rules from text. In *Knowledge Discovery and Data Mining*, pages 323–328.
- Yusuke Shinyama and Satoshi Sekine. 2004. Paraphrase acquisition for information extraction. In *The Second International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP2003) at ACL 2003*. Sapporo, Japan, July.
- D. D. Sleator and D. Temperley. 1993. Parsing English with a link grammar. In *Third International Workshop on Parsing Technologies*.
- Idan Szpektor, Hristo Tanev, Ido Dagan, and Bonaventura Coppola. 2004. Scaling web-based acquisition of entailment relations. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 41–48, Barcelona, Spain, July. Association for Computational Linguistics.
- René Witte and Sabine Bergler. 2003. Fuzzy Coreference Resolution for Summarization. In *Proceedings of 2003 International Symposium on Reference Resolution and Its Applications to Question Answering and Summarization (ARQAS)*, pages 43–50, Venice, Italy, June 23–24. Università Ca’ Foscari.